Fuzzy Entropy and Similarity Measures for Water Quality Assessment

Xiaojing Wang, Zhihong Zou, and Tianyi Fu

***Abstract*—Similarity measures were used to assess the water quality of 7 sites of Tai Lake in 2011. The data were collected once a week and there were 364 samples totally. In each samples, the concentrations of three indicators were collected: DO, COD, NH3-N. For the water quality data was a fuzzy set, semi-normal distribution membership functions were used to obtain the pollution condition of a given sample. The calculation results of fuzzy entropies were used to determine the weights of the indicators. The weight of DO was higher than the other two indicators. This also explained the affection of DO on other indicators. The similarity degree of the results of three similarity measures was 76.10%. This illustrated that they were suitable to determine the water quality ranks. Hamming similarity measures and Euclid similarity measures gave 94.51% of the samples the same water quality ranks. The sample site Jishuigang, Shanghai City was taken to illustrate.**

***Index Terms*—Water quality assessment, similarity measures, fuzzy entropy, entropy weights.**

I. Introduction

Along with the development of economic and science technology, the environment has been polluted seriously. For it is direct connection with our everyday life, water quality has become a focus that the society pays much attention.

Many methods were used to assess the water quality. An index model for quality evaluation of surface water quality classification was proposed using fuzzy logic [1]. Multivariate statistical methods were used to assess the water quality [2]-[7]. Fuzzy matter-element methods were also a useful tool to determine the water quality ranks [8].

The water quality data could be considered as the fuzzy data and be analyzed with fuzzy mathematics. The term fuzzy in the sense seems to have been first introduced by Zadeh [9]. The theory of Fuzzy Sets, a technical exposition of just a mathematics, were also termed by Zadeh [10]. Similarity measures are methods which could evaluate how similar the two samples is.

There are many uses of similarity measures. Adabitabar Firozja [11] developed a distance measure between fuzzy numbers as an interval number and metric was used to define a similarity measure on fuzzy numbers as an interval number. Fuzzy similarity measures were used to analysis the water resources rational allocation [12]. As for the combination forecasting with the interval values, three new combination forecasting models were established based on similarity measures and distance measures and the basic property of these models were studied [13]. A lake water quality evaluation model based on the principle of fuzzy matter-element is put forward in combination of the related characteristics of water evaluation and the similarity measures [14]. Synthetic evaluation based on similarity measures was applied to verify the air pollution [15]. Through the introduction of the biggest-smallest approach degree, a new indicator of the relevance of the optimal combination forecasting model was established [16]. Zwick [17] used the similarity measure based on Minkowski's r-metric. Normalization approach was introduced to the application of similarity measure by Cross [18].

Manuscript received May 7, 2013; revised July 12, 2013.

Xiaojing Wang, Zhihong Zou, and Tianyi Fu are with the School of Economics and Management of Beihang University, Beijing 100191 China (e-mail: star_wxj@126.com, zouzhihong@buaa.edu.cn, sophia_1118@sina.com).

II. Theories and Methods

A. Sample Sites

Lake Tai, or Tai Lake is a large lake in the Yangtze Deltaplain, on the border of Jiangsu and Zhejiang provinces in eastern China. The waters of the lake belong to Jiangsu Province in its entirety with part of its southern shore forming the boundary between the two provinces. With an area of 2,250 km ²and an average depth of 2 meters, it is the second largest freshwater lake in China, after Lake Poyang. The lake has about 90 islands, ranging in size from a few square meters to several square miles.

Lake Tai is linked to the renowned Grand Canal and is the origin water resources of a number of rivers, including Suzhou Creek. In recent years, it has been plagued by pollution as a result of rapid economic growth in the surrounding region.

Pollution of the lake has been ongoing for decades despite efforts to reduce pollution that were not sustained and thus proved ineffective. In the 1980s and 1990s, the number of industries in the lake region has tripled, while the population also increased significantly. One billion tons of wastewater, 450,000 tons of garbage and 880,000 tons of animal waste were dumped in the shallow lake in 1993 alone. The central government intervened and initiated a campaign to clean up the lake and set a deadline to comply with pollution standards. When the deadline was not met, 128 factories were closed on New Year's Eve in 1999. Compliance improved somewhat afterwards, but the pollution problem remained severe. In May 2007, the lake was overtaken by a major algae bloom and major pollution with cyanobacteria. The government called the lake a major natural disaster despite the anthropogenic origin of this environmental catastrophe. The lake provides water to 30 million residents, including about one million in Wuxi. By October 2007, it was reported that

the Chinese government had shut down or given notice to over 1,300 factories around the lake. However, in 2010 not a single factory was closed and the Economist reported that a fresh pollution outbreak had occurred in this year.

B. Membership Functions

For different water quality ranks, the membership functions are usually different. They are used to measure the degree that a given sample belongs to a rank. And each water quality rank has a group of membership functions. This cannot make a consistency of the water quality. Semi-normal distribution membership functions were used because these measures gave the different levels of indicators a unified measurement criterion.

For indicators, such as DO (dissolved oxygen) when the value is larger, the pollution is lighter. In this situation, the increasing semi-normal distribution has been used. The calculation formula can be described in Eq. 1.

$$\mu_{ij}(X_{ij}) = \begin{cases} 0 & X_{ij} \le Y_{j5} \\ 1 - e^{-(\frac{X_{ij}-Y_{j5}}{\sigma})^2} & X_{ij} > Y_{j5} \end{cases} \quad (1)$$

And we can see that, below the boundary value of the fifth rank, the membership function is 0. This means that the pollution of a water system is very serious because of lack of DO and some measures should be taken to deal with the situation. When the pollution is not more serious than fifth rank, the membership function can be expressed with the mathematical formula: $1 - e^{-(\frac{X_{ij}-Y_{j5}}{\sigma})^2}$, where σ is the standard deviation of the boundaries of standard water quality ranks (Fig. 1).

For other indicators, such as COD, NH3-N, confirmation of membership function uses decreasing semi-normal distribution. The calculation formula is described in Eq. 2.

$$\mu_{ij}(X_{ij}) = \begin{cases} 1 & X_{ij} \le Y_{j1} \\ e^{-(\frac{X_{ij}-Y_{j1}}{\sigma})^2} & X_{ij} > Y_{j1} \end{cases} \quad (2)$$

The membership function has the similar means with the increasing semi-normal distribution from the opposite direction. Because the pollution is more serious when the value is larger, the value of membership function is smaller and smaller along with the concentration becomes larger and larger. The value of member function is one when the concentration is less than the boundary of rank one.

C. Fuzzy Entropy and Weights

Many methods are used to determine the weights of water quality indicators. The two main kind methods are from objective and subjective aspects. The weights determined by subjective methods are according to the experts' experienced judgment. Also it is much more developed, it is used less for different experts holding different opinions. Objective weights determined methods are conducted based on the actual data and don't depend on subjective judgment, such as coefficient of variation. Coefficient of variation obtains the weights of indicators according to handling of the observed value. Another method is to use entropy to determine the weights.

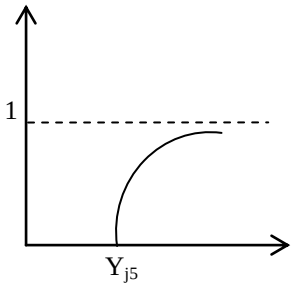

Fig. 1. The increasing semi-normal distribution function.

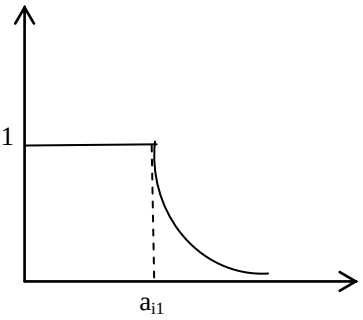

Fig. 2. The decreasing semi-normal distribution function.

Entropy is the measure of the amount of information that is missing before reception. It can be used to measure the amount of useful information with the data provided. It was first applied in thermodynamics and was introduced into the information theory [19]. The definition of the entropy is expressed in terms of a discrete set of probabilities $p(x_i)$: $H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i)$. To determine the weights of indicators with entropy is an objective method. Entropy weight method calculates the entropy weight of each indicator according to the variation of each indicator. Then revising the weight of each indicator and obtaining the objective results. The smaller the entropy of the indicator is, the larger the variation of the indicator. This kind of indicators plays an important role and the weights of them are bigger. The larger the entropy of the indicator is, the smaller the variation of the indicator. The weights of this kind of indicators are smaller.

In this study, fuzzy entropy method is used to determine the entropy of indicators and to calculate the weights based on the entropy. This method is suitable for the situation where the data are fuzzy sets.

The entropy of a fuzzy subset, X, of the finite set $\{x_1, x_2, \cdots, x_n\}$ with respect to a probability distribution $P = \{p_1, p_2, \cdots, p_n\}$ can be defined as follows: $H^F(A) = -\sum_{i=1}^{n} \mu_A(x_{ij}) f_{ij} \log f_{ij},\ i = 1, 2, \cdots, m$. In which $f_{ij} = (r_{ij} + 1) / \sum_{j=1}^{n} (r_{ij} + 1)$ and $r_{ij} = \sum_{j=1}^{n} (\mu_{ij} + 1)$.

μ_A is the membership function of fuzzy set A. The equation represents the entropy of a fuzzy event A with respect to the membership function μ_A.The weight of the entropy for the ith indicator could be defined as $\omega_i = (1 - H_i) / (m - \sum_{i=1}^{m} H_i)$ in which $0 \le \omega_i \le 1$, $\sum_{i=1}^{m} \omega_i = 1$.

D. Similarity Measure

When calculating the similarity measures, some expressions should be defined first. Supposed that *X*, *Y* are two fuzzy sets, there are some formulas used to express the operation relationships. The result of the expression $\mu_X(X_{ij}) \wedge \mu_Y(Y_{ij})$ describes the minimum value of two fuzzy membership functions $\mu_X(X_{ij})$ and $\mu_Y(Y_{ij})$. While the expression $\mu_X(X_{ij}) \vee \mu_Y(Y_{ij})$ shows the maximum of the two fuzzy membership functions. These two expressions are the basic operations of fuzzy mathematics and are used widely.

There are many kinds of similarity measures. Three kinds are used to assess the water quality of Tai Lake. They are Hamming similarity measures, Euclid similarity measures and max-min similarity measures. The equations of them are shown in Eq. 3 to Eq. 5.

Hamming similarity measures

$$N_H(X_i, Y_j) = 1 - \frac{1}{n} \sum_{k=1}^{n} \left| \omega_k (\mu_X(X_{ik}) - \mu_Y(Y_{jk})) \right| \quad (3)$$

Euclid similarity measures

$$N_E(X_i, Y_j) = 1 - \sqrt{\frac{1}{n} \sum_{k=1}^{n} [\omega_k (\mu_X(X_{ik}) - \mu_Y(Y_{jk}))]^2} \quad (4)$$

Max-min similarity measures

$$N_M(X_i, Y_j) = \frac{[\sum_{k=1}^{n} \omega_k (\mu_X(X_{ik}) \wedge \mu_Y(Y_{jk}))]}{[\sum_{k=1}^{n} \omega_k (\mu_X(X_{ik}) \vee \mu_Y(Y_{jk}))]} \quad (5)$$

The means of similarity measures and distance are opposite. If the similarity measures are close to 1, this illustrates that the two fuzzy sets are more similar. If the similarity measures are close to 0, the difference between two fuzzy sets is much greater.

III. Results and Discussion

For Tai Lake, three water quality indicators of 7 sample sites were collected. In each sample site, the concentrations of three indicators (NH3-N, DO, COD) were obtained for 52 weeks in 2011. So there were 364 samples totally. The Chinese official water quality ranks were also given and they were obtained according to the worst indicator among the ranks of the indicators. The original monitoring data were recorded as matric X.

TABLE I: Boundary Values of Some Indicators in Water Quality Standards (GB3838-2002)

Indicator	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
DO (mg/L)≤	7.5	6	5	3	2
COD (mg/L) ≤	15	15	20	30	40
NH3-N (mg/L) ≤	0.15	0.5	1	1.5	2

TABLE II: Entropy and Weights of the Indicators

Indicators	DO	COD	NH3-N
Entropy	122.94	47.49	75.40
Weight	0.50	0.19	0.31

A. Determining of Entropy and Weights

The membership functions of three indicators for the 364 samples were calculated by the semi-normal distribution functions. The σ for each indicator was determined by the standard deviation of the boundaries of standard water quality ranks (Table I). The boundary values of three indicators in water quality standards (GB3838-2002) were recorded as matric Y. The membership functions of three indicators for the boundaries of the standard water quality ranks also need to be calculated.

When the membership functions were calculated, the fuzzy entropy could be obtained according to the membership functions and the formula $H_i^F = -\sum_{j=1}^{5} \mu(Y_{ij}) f_{ij} \log f_{ij},\ i = 1, 2, 3$, where $f_{ij} = (r_{ij} + 1) / r_i$ and $r_i = \sum_{j=1}^{5} (\mu_{ij} + 1)$. μ_{ij} was the membership function of the fuzzy set. The equation represents entropy of fuzzy event with respect to membership function μ. The weight of the ith indicator could be defined as $\omega_i = (1 - H_i) / (m - \sum_{i=1}^{m} H_i)$. The weights of the indicators have constraint: $0 \le \omega_i \le 1$, $\sum_{i=1}^{m} \omega_i = 1$.

From Table II, we could draw the conclusion that the weight of DO was 0.50. This meant that when assessing the water quality, DO played a crucial rule and the concentration of it affected the water quality ranks more than other indicators. This was because the concentrations of COD and NH3-N were affected when the concentration of DO changed. NH3-N can be oxidized by DO and the COD (chemical oxygen demand) were determined by the DO (dissolved oxygen). This result was consistent with the actual situation.

B. Results Analysis

The water quality ranks calculated by the similarity measures were little higher than the official results because the water quality ranks calculated by similarity measures took into account all of the three indicators. When the three indicators were all considered, the role of the worst indicators was weakened and the ranks of the samples must be higher than the official results.

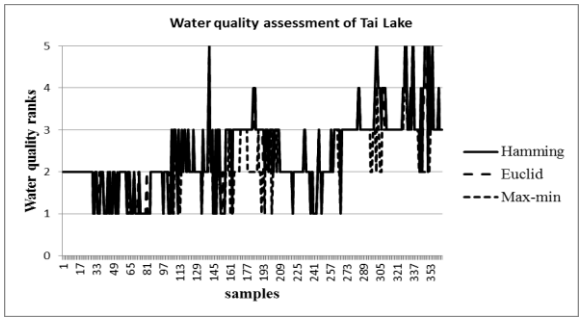

Fig. 3. Results of the three similarity measures for the 364 samples.

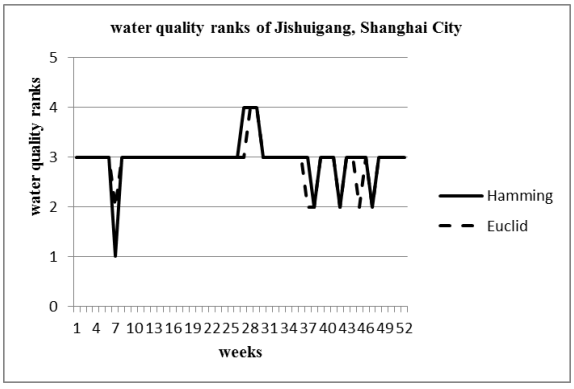

Fig. 4. Results of the three similarity measures for Jishuigang, Shanghai City.

The results of three similarity measures had some differences (Fig. 3). For most of the cases, the ranks determined by the three methods were very similar. A specific sample could be classified into the proper water quality ranks with different methods. Among the 364 samples, there were only 23.90% (87 samples) that were not classified into the same water quality ranks. This meant that the similarity measures methods were stable to assess the water quality of Tai Lake. Though the similar percent was enough to illustrate the availability of the methods mentioned above, the results of the three also had some differentiation. They are shown as below.

When only considering Hamming similarity measures and Euclid similarity measures, 94.51% of the samples (344 samples) were classified into the same water quality ranks. The calculations of these two methods were similar, because they all used the difference between the membership functions of samples and the membership functions of water quality ranks.

But the similar degree of Hamming and max-min similarity measures was only 78.57%. This differentiation may be caused by the information losing of the max-min similarity measure. Form the Eq. 5, it was obviously that during the calculation of max-min similarity measure, not only the procedure of $\mu_X(X_{ij}) \wedge \mu_Y(Y_{ij})$ but also the procedure $\mu_X(X_{ij}) \vee \mu_Y(Y_{ij})$ omitted one side of two membership functions. This made the information lost and the results were not accurately enough. Similar results were also obtained when comparing the Euclid with max-min similarity measures and the similar degree was 76.92%.

The assessment results also demonstrated the seasonal trend in each site. Jishuigang of Shanghai City was taken as an example to illustrate the assessment results of Hamming and Euclid similarity measures, for the results of these two methods were more similar.

From Fig. 4, we could find that the water quality ranks became to change from week 30. Before this time, the water quality ranks were about 2 ranks and sometimes it was 3 ranks for some uncontrolled factors. But after week 30 of 2011, the water quality ranks had fluctuation more than usual. This because during that season, there were more rainfall and this could dilute the concentration of the pollutions and improve the water quality.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 51178018, 71031001).

REFERENCES

[1] I. Yilmaz, "Fuzzy evaluation of water quality classification," *Ecological Indicators*, vol. 7, pp. 710-718, March 2007.
[2] F. I. Cansu, E. Ozgur, I Semra, A. Naime, Y. Veysel, and A. Seyhan, "Application of multivariate statistical techniques in the assessment of surface water quality in Uluabat Lake, Turky," *Environental Monitoring and Assessment*, vol. 144, pp. 269-276, 2008.
[3] P. Agelos, M. Athina, H. Christos, P. Panagiotis, P. Olga, and D. Eleni, R. Ioannis, "Application of multivariate statistical methods for groundwater physicochemical and biological quality assessment in the context of public health," *Environental Monitoring and Assessment*, vol. 170, pp. 87–97, 2010.
[4] K. S. Sanjay, "Application of multivariate statistical techniques in hydrogeochemical studies—a case study: Brahmani–Koel River (India)," *Environental Monitoring and Assessment*, vol. 164, pp. 297–310, 2010.
[5] K. P. Singh, A. Malik, and S. Sinha, "Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques—a case study," *Analytica Chimica Acta*, vol. 538, pp. 355–374, 2005.
[6] P. S. Kunwar, M. Amrita, M. Dinesh, and S. Sarita, "Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomi River (India)-A case study," *Water Research*, vol. 38, pp. 3980-3992, 2004.
[7] P. S. Kunwar, M. Amrita, and S. Sarita, "Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques-A case study," *Analytica Chimica Acta*, vol. 538, pp. 355-374, 2005.
[8] D. J. Liu and Z. H. Zou, "Water quality evaluation based on improved fuzzy matter-element method," *Journal of Environmental Sciences*, vol. 24, pp. 1210-1216, July 2012
[9] L. A. Zadeh, "From circuit theory to system theory," in *Proc. the IRE*, vol. 50, 1962, pp. 856-865.
[10] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338-353, May 1965.
[11] M. A. Firozja, G. H. Fath-Tabar, and Z. Eslampia. "The similarity measure of generalized fuzzy numbers based on interval distance," *Applied Mathematics Letters 25*, pp. 1528-1534, 2012.
[12] Y. D. Liu, "Analysis on water resources rational allocation based on fuzzy close," M. S. thesis, Yangzhou University, Yangzhou, China, 2008.
[13] L. Yang, "The interval combination forecasting based on the approach degree and distance measure," M. S. thesis, Anhui University, Hefei, China, 2012.
[14] B. Wu, "Study on fuzzy matter-element based model for lake water quality evaluation," *Water resources and Hydropower Engineering*, vol. 38, pp. 12-15, April 2007.
[15] Y. W. Tong, Z. B. Liu, and H. Chang, "Air environmental quality assessment with similarity measure," *Energy and Environment*, vol. 3, pp. 115-116, 2008
[16] H. J. Yuan and G. Y. Yang, "The combination forecast model based on the biggest-smallest approach degree," *Operations research and management science*, vol. 19, pp. 116-122, February 2010.
[17] R. Zwick, E. Carlstein, and D. Budescu, "Measures of similarity among fuzzy concepts, a comparative analysis," *International Journal of Approximate Reasoning*, vol. 1, pp. 221–242, February 1987.
[18] V. V. Cross and T. A. Sudkamp, *Similarity and compatibility in fuzzy set theory: assessment and applications*, Heidelberg, New York: Physica-Verlag, 2002.

[19] C. E. Shannon, "A mathematical theory of communications," *Bell Systems Technical Journal*, vol. 27, pp. 379-423, March 1948.

Xiaojing Wang was born on February 4, 1984, Hebei Province, China. Wang got her bachelor's degree in management, 2008, Yanshan University, Qinhuangdao, China and master's degree in management, 2011, Beihang Unviersity, Beijing, China.

She is a PH.D candidate in Beihang University, Beijing, China. Now she is interesting in water quality assessment, forecast and water quality models. Doctor Wang obtained the outstanding student leaders and Guanghua scholarship, 2012.

Zhihong Zou was born on May 7, 1958, Heilongjiang Province, China. She received her Ph.D in management engineering from Beihang University, China in 1996.

She is a professor of Beihang University, Beijing, China. Her research interests include water quality assessment and forecast, system simulation and its application. She has hosted 4 projects supported by National Natural Science Foundation of China (NSFC).

Tianyi Fu was born on November 18, 1989, Liaoning Province, China. Fu got her bachelor's degree in management, 2011, Beijing University of Chemical Technology, Beijing, China.

She is a master candidate in Beihang University, Beijing, China. Now she is interested in water quality assessment and forecast. Fu acquired the National Scholarship and Guanghua Scholarship, in 2009 and 2012 respectively.